Hadoop and Spark Technology

- A Comparative Study of Big Data Tools

Rupesh Jaiswal, Akhilesh A. Waoo
Ph.D. Research Scholar, HOD, Department of CS/IT,
AKS University, Satna(MP), India.

Abstract: Data In this Digital world everything is going to be digitalized. People are using social media, digital documents, Online Banking, Facebook, Twitter, etc. All these things generate a bulk amount of data either it is in the structured or unstructured form. To manage this bulk amount of data Hadoop and Spark technologies are used. These technologies are capable to manage, store, process and analyze such kind of data. This paper highlights the overview of Hadoop, MapReduce, HDFS and Apache Spark tools of Big Data.

I. Introduction

Big data can be defined as a bulk amount of Data or various sets of information that is increasing rapidly. Big Data is also data but with a huge size. This vast amount of data can’t be stored and processed by the normal Data Base Management System tools. Big Data considers the six ‘V’s: Volume, Variety, Velocity, Veracity, Variability and the most importantly Value.

Volume: Big Data operates on a bulk amount of data means it is related to a size that is enormous.

Variety: Big Data supports different types of data either it is Structured or Non-structured by nature. Versatile types of data can be stored, accessed, operated and analyzed by Big Data.

Velocity: Velocity relates to speed i.e. a variety of data of large volume can be created and collected very fast.

Veracity: Veracity relates to the true worthiness of the data.

Variability: Variability relates to the contradiction. Different types of data in different volume can affect the process. Big data handles this without any discrepancies.

Value: Value is a very important factor. Data is useless if it is not valuable [1].

II. Keywords

Big Data, Hadoop, HDFS, MapReduce, YARN, Apache Spark.

III. hadoop

Hadoop is an open–source Software framework, written in Java that supports a bulk amount of data storage and processing on a cluster that encloses many hardware systems [2].It has the capability to control virtually limitless coexisting responsibilities or tasks. Hadoop has these main components: HDFS, MapReduce and YARN.

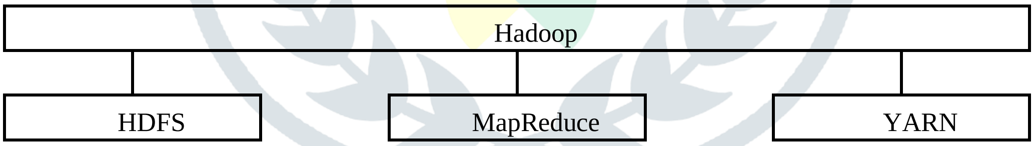

Figure: 1 (Components of Hadoop)

III.1 : HDFS

HDFS(Hadoop Distributed File System) is liable for storing data of different formats transversely the cluster. HDFS provides the facility to access the data very fast. HDFS also provides the Parallel Analysis system.[7]

HDFS has two main Nodes:

a) NameNode and
b) DataNode.

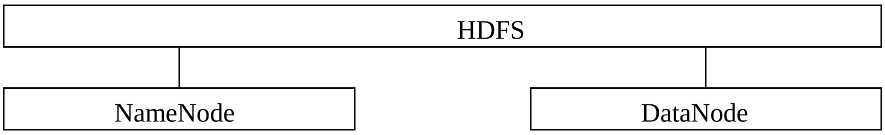

Figure: 2 (Nodes of HDFS)

III.1.1: NameNode:

NameNode is a Master Node that provides information about the data which is used in DataNode. In other words it stores MetaData.

III.1.2: DataNode:

DataNode is a Slave Node which stores the original and actual data.

III.2: MapReduce:

Map-Reduce is a software framework which is used to write applications and parallel processing of vast amount of data which is stored on large clusters of commodity hardware in a reliable as well as fault-tolerant manner [3]. Map and Reduce are the two essential tasks of the MapReduce algorithm. The main principle of the Map task is to take a large set of data and convert it into another set of data that is broken down into distinct tuples or Key/Value pairs. Next the output of Map task, the tuple which constructs the input for a reduction task. The data tuples are then reduced and converted into a smaller set of tuples[6]. Always after the Map task the Reduce task is executed. The major strength of the Map-Reduce framework is scalability. A Map-Reduce program can easily be used to work over a cluster which has hundreds of nodes or even thousands of nodes.[5]

III.3: YARN

The second component of Hadoop is YARN.(Yet Another Resource Negotiator). YARN is responsible for resource management. YARN handles the allocation of resources and controls entire processing activities by scheduling the tasks. ResourceManager and NodeManager are the two main parts of YARN.

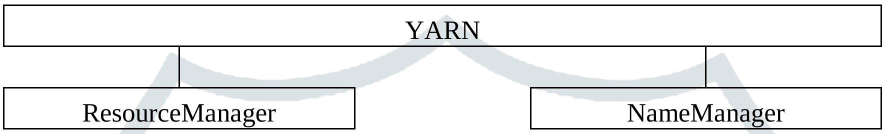

Figure: 3 (Parts of YARN)

III.3.1: ResourceManager

ResourceManager is the first main component of YARN. The main function of ResourceManager is to accept the processing requests and send them to the appropriate NameManager.

III.3.2: NameManager

NameManager is the second main component of YARN. The main function of NameManager is to execute the processes on every DataNode.

IV. Apache Spark

Spark is a project of Apache and it is famous as “lightning-fast cluster computing”. Spark is open-source and it is an Engine that can process a broad range of data. Spark works faster than Hadoop technology. It provides a hundred times faster in memory and ten times faster on disk, speed than Hadoop. Spark writes the codes very speedily as more than eighty operators can be used in it. This reduces the number of code lines.

Spark provides APIs in Scala, Java, Python and R [4]. And it can be integrated with Hadoop EcoSystem, HDFS, etc. and can work standalone also.

Spark can do batch processing as well as stream processing. Batch processing relates to the processing of the formerly composed job in a single batch, whereas stream processing can be defined to deal with Spark streaming data.

Spark can be executed on Hadoop Clusters. It is not the extension of Hadoop; but it can extent Hadoop MapReduce to the higher stage. Spark also includes iterative queries and stream-processing. It combines all the Big Data tools. Spark has a self cluster management system. Spark uses Hadoop for storage purposes only.[8]

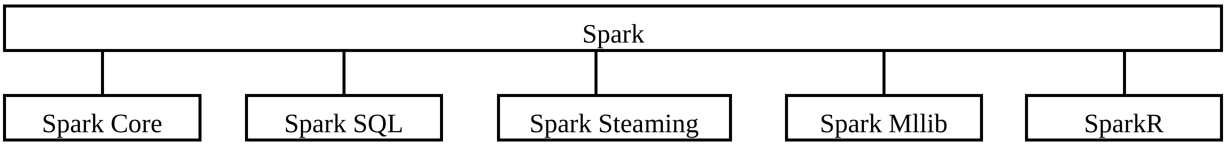

Figure: 4 (Parts of Spark)

V. Discussion:

This paper presents some Big Data tools like Hadoop, HDFS, MapReduce, YARN and Apache Spark and learned how these tools are more capable than other technologies.

VI. Conclusion:

To conclude, firstly Hadoop has a powerful distributed storage file (HDFS), with performances and features that are available for Big Data processing systems such as Hadoop MapReduce and Apache Spark. Hadoop MapReduce is a platform built for batch processing. It was very famous and very usable by multiple companies in several areas because Hadoop MapReduce was the first technology that analyzes mass distributed data.

On the other hand, Apache Spark is the new brightest platform on Big Data technology. It has better performances. The Spark is very profitable thanks to the data processing in memory. It is well-matched with all of the data sources & file formats of Hadoop, also it has several APIs. Apache Spark even includes graph processing and machine-learning capabilities. Although some companies feel compelled to choose between Hadoop and Spark, the fact that isn't a direct fight. They are complementary technologies that work in tandem in some cases, or separately in other cases, depending on the data or the desired results. The truth is that Spark and Hadoop have a dependent relationship with each other. Hadoop provides options that Spark doesn't possess as HDFS. Spark provides real-time in-memory processing. The perfect scenario of Big Data is that Hadoop and Spark work together on the same team.

VII. ACKNOWLEDGEMENT:

I state my special thanks and regards to my Ph.D. guide Mr. Akhilesh A. Waoo, HOD AKS University for his encouragement, support and Guidance.

REFERENCES:

[1] Saman Sarraf, Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, L8S 4L8, Canada Rotman Research Institute at Baycrest, University of Toronto.

[2] Houssam BENBRAHIM, Hanaa HACHIMI and Aouatif AMINE, Proceedings of the International Conference on Industrial Engineering and Operations Management Bangkok, Thailand, March 5-7, 2019, Comparison between Hadoop and Spark, GS Laboratory "LGS", BOSS-Team. National School of Applied Sciences, Ibn Tofail University, Kenitra, Morocco.

[3] Comparative analysis of Hadoop tools and Spark technology, Aniket Wakde, Purvesh Shende, Sudarshan Waydande, ShravaniUttarvar Ganesh Deshmukh, Computer Department, Pimpri Chinchwad college of Engineering, Pune, India

[4] [Online] Available https://spark. apache.org

[5] Can Uzunkva, Tolga Ensari, Yusuf Kavurucu, "Hadoop Ecosystem and Its Analysis on Tweets" in World Conference on Technology, Innovation and Entrepreneurship.

[6] Jeffrey Dean and Sanjav Ghemawat. "MapReduce: Simplified Data Processing on Large Clusters".

[7] http://www.edureka.co definition and introduction of HDFS.

[8] David Andre'sic, Petr Saloun and loannis Anagnostopoulos- " Efficient Big Data Analysis on a Single Machine using Apache Spark and Self Organizing Map Libraries".